UNLIMITED

A238 894

RSRE

MEMORANDUM No. 4463

ROYAL SIGNALS & RADAR ESTABLISHMENT

ANALYSIS OF THE PHONEME RECOGNITION PERFORMANCE OF THE ARM CONTINUOUS SPEECH RECOGNITION SYSTEM

Author: S R Browning

PROCUREMENT EXECUTIVE,
MINISTRY OF DEFENCE,
RSRE MALVERN,
WORCS.

RSRE MEMORANDUM No. 4463

91-06119

UNLIMITED

0100492

CONDITIONS OF RELEASE

302014

** DRIC U

** DRIC Y

Reports quoted are not necessarily available to members of the public or to commercial organisations.

Royal Signals and Radar Establishment

Memorandum 4463

Analysis of the phoneme recognition performance of the ARM continuous speech recognition system

S R Browning

21st March 1991

Abstract

This memorandum presents the results of a phonetically motivated analysis of the speech recognition system developed as part of the *ARM* (Airborne Reconnaissance Mission) project. The aim of the work described here is to investigate to what extent errors can be explained by phonetic effects; those which cannot may indicate where models may be improved. The background to the investigation, and the problems of evaluating phoneme recognition performance are described, then the remainder of the report is concerned with a detailed analysis of specific types of errors, motivated by a desire to find phonetic explanations of them.

INTENTIONALLY BLANK

Contents

List of tables

List of figures

1. Introduction

This memorandum presents the results of a phonetically motivated analysis of the speech recognition system developed as part of the *ARM* (Airborne Reconnaissance Mission) project. The aim of the *ARM* project is accurate recognition of continuously spoken airborne reconnaissance reports using sub-word (phoneme) hidden Markov modelling techniques. The version of the system on which this study is based is speaker-dependent and has a vocabulary of 497 words. The *ARM* system is described in [5]. The version of the system on which this investigation was based scores an average of 86.8% word accuracy with word level syntax (i.e. perplexity = 497).

The aim of the work described here is to investigate to what extent errors can be explained by phonetic effects; those which cannot may indicate where models may be improved. For instance, if /p/ is misrecognised as /b/, this is understandable from the phonetic point of view as the two are acoustically rather similar; however, if /p/ were to be consistently misrecognised as /@U/ or /z/ this error would be difficult to explain in acoustic-phonetic terms, and would probably indicate that there is something wrong with the model(s).

The following section describes the background to the investigation, and the problems of evaluating phoneme recognition performance. The remainder of the report is then concerned with a detailed analysis of specific types of errors, motivated by a desire to find phonetic explanations of them. The phonemic transcriptions in this report are in the SAM-PA notation [2, 8], and see Appendix A for the list of phonemes and examples.

2. Background

2.1 The *ARM* task

The airborne reconnaissance mission reports which the *ARM* system recognises follow a standard format, beginning with some highly structured sentences recounting the mission details, such as time and place of observation. Then follows a slightly more free-format section where the reconnaissance pilot describes what he sees and assesses its condition. The report concludes with a brief description of the weather and visibility conditions. The vocabulary of the system with its citation-form phoneme transcription can be found in Appendix B. An example of an *ARM* report is given below.

> *Recce report two stroke charlie stroke six eight one. Military activity at map co-ordinates india hotel eight four three four. Time over target eleven oh seven GMT. New target cat zero one; operational airstrip. Roughly fifteen light aircraft of type possibly foxbat. Main runways heading southwest wholly unusable, SAM defences intact. TARWI fife eighths at niner hundred; end of report.*

It is important to note that this is not a natural use of language, and this may influence the generality of the results of this study, in that the relative frequency of phonemes in the *ARM* vocabulary will not necessarily match that in natural language. In particular, the phoneme /D/, which ranks eighth in normal use (due to the high frequency of the word "the" in natural speech), does not occur at all in the data of two of the speakers examined here, and only occurs once for Speaker 2, who pronounces the word "with" as /wID/, rather than /wIT/. A comparison of phoneme frequencies in normal speech [3] with those in the *ARM* data can be found in Appendix A.

2.2 The Speakers

The system currently recognises the speech of three speakers, and is trained separately for each, using approximately fifteen minutes of speech (airborne reconnaissance mission reports) from each of the three. Speakers 1 and 2 are male; Speaker 1 is basically RP, while Speaker 2 has Midlands overtones. Speaker 3 is female and has north-eastern colours in her accent. Each speaker has their own dictionary to take account of dialectal variations. In this report I will be trying to draw some general conclusions about error types which apply to all three speakers, but the more important speaker differences will also be pointed out.

2.3 The System

The *ARM* system is described in detail in [7]. Sub-word (phoneme-like) hidden Markov models are used, but it is well known that the acoustic realisation of phonemes varies in different contexts. In order to take account of this context-sensitivity, approximately 1500 triphones are used. Triphone modelling assumes that it is the immediately surrounding context which exerts the most influence on the acoustic realisation of a particular phoneme, so a triphone is a model of a phoneme in its left and right context. In the current system this is restricted to word-internal contexts. See [6] for a full description of the triphone methods used in the *ARM* system.

In addition to the triphones for each context-sensitive phoneme, a number of short words are modelled explicitly at the word level. Non-speech sounds, such as breath noise or lip smacks are also modelled explicitly with a set of single state models. Both the word models and the noise models are treated in exactly the same way as the triphones.

For the purposes of the analysis described here the system was configured as a phoneme recogniser, with no dictionary and no syntax. There is, however, some measure of constraint, in that the right context of each triphone must match the left context of the next. This is no small constraint; as the triphones are word-internal, and the vocabulary so limited, the number of different triphones that actually occur is very small. (There are 1456 different triphones in the *ARM* set, while a 68000-word dictionary has 14378.)

2.4 Evaluating the performance

The system has so far been tested on ten *ARM* reports (that were not in the training set) from each speaker, containing a total of approximately 2290 phonemes per speaker, 6873 in all. The arrangement described above produced an average phoneme error rate of 26.2% for the three speakers. Phoneme recognition performance is measured by aligning the output of the system with a phonemic transcription of the test material. The latter is obtained by replacing each word in the orthographic transcription of the data with its phonemic transcription from the (speaker-dependent) dictionary. Errors are classified as substitutions, deletions or insertions. Substitutions occur when a phoneme is misrecognised as another phoneme, deletions when a phoneme has been missed by the system, and insertions when the system has recognised an extra phoneme. Recognition performance is stated in terms of correctness and accuracy. The first is simply a measure of how many times the system produced the same label as the dictionary transcription, while the second is a more stringent measure, which is calculated by subtracting the number of insertions from the number of correctly recognised phonemes, and as such is a more satisfactory indicator of the recognition performance.

The alignment of recognition results and transcription is automatic, and a summary of individual phoneme performance is also produced, along with a confusion matrix. However, this process is not accurate, in that sometimes errors in the alignment obscure correct matches, and insertions are counted as substitutions, etc. If alignment errors are taken into account the overall results are not significantly different (on average slightly over 1% either way), but as it is the distribution and details of the error types that are most important for this investigation, it is necessary to hand correct the alignment and scoring. This is quite a lengthy process as it involves listening to the speech at the same time as observing the labelling produced by the system on the spectrograms, and then re-compiling the phoneme statistics and confusion matrix. All the analyses in this paper are based on hand-corrected alignments.

It is in practice extremely difficult to assess performance, as in many cases the speaker may not actually produce the somewhat idealised pronunciation represented in the dictionary. For example, in the sequence "six six" the speaker is likely to produce only one /s/ (though it may be somewhat lengthened) for the two which phonemically occur over the word boundary. In this example, if the system recognises only one /s/ it is penalised for having deleted a phoneme. There are numerous examples of this nature, and these will be discussed under the appropriate categories below. In spite of these shortcomings, results are scored strictly against the dictionary transcription in order to ensure that the evaluation system is both consistent and automatic. We are, however, currently investigating the inclusion of alternative transcriptions in the dictionary, which will allow us to take account of many of these so-called errors.

3. The Analysis

In this section the phoneme recognition results are analysed in some detail. A summary of the phoneme recognition results for each speaker and for all speakers combined is shown in Table 1. From this it can be seen that the results for all three speakers are in the same range, although Speaker 2 and Speaker 3 have slightly better performance than Speaker 1. This general trend is evident in most of the more detailed analyses of phoneme performance; particular differences between speakers will be pointed out below.

Speaker	% correct	% substitution	% deletion	% accuracy	total no. of phonemes
1	75.5	13.2	11.3	68.8	2290
2	82.0	10.9	7.1	76.8	2290
3	80.8	11.6	7.6	76.0	2293
All speakers	79.4	11.9	8.7	73.8	6873

Table 1 Summary of phoneme recognition results

The rest of Section 3 is devoted to a discussion of the different types of error. A complete set of tables showing the individual phoneme performance for each speaker and all speakers combined can be found in Appendix C (tables C1–4). For convenience and clarity in the following sections only the information about the factors under discussion will be presented.

3.1 Analysis of Correctness/accuracy

3.1.1 Individual phonemes

Table 2 shows the phoneme correct and accuracy scores for each speaker and all speakers. A number of phonemes (/D, oI, 3, e@ and U/) occur so rarely in the *ARM* reports that their results are unreliable indicators of performance, so these will be ignored in this analysis. The models with the poorest performance were those for whole words, which tended to be confused with one or more phonemes. Although there are slight differences between speakers at the top end of performance, it can be seen that in general /A/ was recognised most reliably, closely followed by /eI/, /S/ and /O/.

It is at the lower end, however, that more obvious differences emerge. None of the speakers has good performance for /N/, for instance, and only Speaker 2 has a reasonable score for /v/. In Speaker 1's data, /p/, although well recognised, suffered from some insertions, and was one of the least accurately recognised phonemes. For this speaker too /m/ scored particularly badly, but /d/ was the least accurate due to an unusually large number of insertions (these will be discussed later). It was /V/ that was least correctly recognised in Speaker 2's case, and this phoneme was relatively often inserted too, making it the least accurate phoneme for this speaker. This may be due to there being relatively few occurrences of this phoneme in Speaker 2's data (18 as opposed 35 in the other speakers' reports. In Speaker 3's case the least correct (apart from /v/ and /N/ which was common to the other speakers) was /b/.

In trying to find general trends in phoneme recognition performance the phonemes were classified into phonetically motivated groups, namely 'manner' and 'place' of articulation. (I have disregarded the word–level models in this classification.) Under 'manner' there is a broad classification into vowels and consonants, which should be self–evident, and a finer one where consonants are split into more specific classes. A list of the members of each of these classes is given in Figure 1, along with the total number of phonemes in each class. The fineness of the place classification was chosen in an attempt to make sure that there were enough members of a given class to give a reasonable sample. In the case of the centring diphthongs and palatal–alveolars there are probably too few, but it would not have been reasonable to include them in any of the other classes. It would therefore be unwise to draw any conclusions about these two classes. It may be useful to note that in general these classes of phonemes occur comparatively rarely, either in normal speech or in the *ARM* test data (see Appendix A). So /e@/ and /I@/ rank 40th and 41st in normal speech and 40th and 30th in *ARM*. And for the palatal–alveolars, /S/ ranks 31st in normal, 27th in *ARM*; /tS/ 38th (37th); /dZ/ 36th in both; /j/ ranks 32nd in normal, 33rd in *ARM*, and /Z/ does not occur at all in *ARM*, and ranks 43rd in normal.

	Speaker 1		Speaker 2		Speaker 3		All		
Phon	% Cor	% Acc	% Cor	% Acc	% Cor	% Acc	% Cor	% Acc	Total
s	83.8	75.7	86.8	80.9	86.8	83.8	85.8	80.2	408
z	86.0	70.2	71.9	50.8	82.5	77.2	80.1	65.5	171
S	83.9	83.9	100.0	100.0	96.8	96.8	93.5	93.5	93
f	89.8	85.9	89.8	83.3	83.4	80.8	87.6	83.3	234
v	36.7	20.0	83.3	70.0	50.0	36.7	56.7	42.3	90
T	83.3	72.2	65.7	54.3	75.0	69.4	74.8	65.5	107
D	-.-	-.-	0.0	0.0	-.-	-.-	0.0	0.0	1
h	61.5	61.5	76.9	76.9	61.5	46.2	66.7	61.6	39
tS	66.7	66.7	88.9	88.9	88.9	88.9	81.5	81.5	27
dZ	72.7	72.7	81.8	63.6	63.6	63.6	72.7	66.6	33
p	92.7	58.5	90.2	61.0	82.9	58.5	88.6	59.3	123
b	60.0	46.7	60.0	60.0	53.4	26.7	57.9	44.4	45
t	85.3	81.8	85.3	83.1	87.0	82.7	85.9	82.6	693
d	59.8	19.5	63.4	52.4	70.0	59.8	64.6	43.9	246
k	89.7	84.5	92.7	88.7	90.7	89.7	91.1	87.6	291
g	75.6	75.6	78.0	78.0	87.8	87.8	80.5	80.5	123
m	38.8	34.7	85.7	71.4	91.8	85.7	72.2	64.0	147
n	55.0	54.4	80.1	73.7	77.2	76.0	70.8	66.7	513
N	55.6	55.6	50.0	44.4	50.0	50.0	51.9	50.0	54
l	78.7	70.7	86.7	81.3	76.0	66.7	80.5	72.9	225
r	88.4	84.5	96.1	96.1	88.4	85.4	90.9	88.7	309
w	79.6	79.6	86.4	81.8	90.9	86.4	85.6	82.6	132
j	100.0	85.7	71.4	71.4	85.7	85.7	85.7	80.9	42
i	84.1	77.3	84.1	80.6	93.7	93.7	86.1	82.1	224
I	65.5	53.8	70.9	70.1	76.3	66.9	72.1	64.0	394
E	74.7	72.3	84.3	83.1	85.6	80.7	81.5	78.7	249
{	90.2	88.2	89.5	84.2	86.0	86.0	88.5	86.1	165
A	97.6	97.6	100.0	97.1	94.3	91.4	97.3	95.5	111
Q	76.9	76.9	96.7	96.7	84.6	84.6	89.3	89.3	56
O	81.1	78.4	94.6	94.6	100.0	100.0	91.9	91.0	111
U	100.0	100.0	66.7	33.3	66.7	66.7	77.8	66.7	9
u	72.7	67.3	83.6	81.8	69.1	67.3	75.2	72.2	165
3	100.0	66.7	100.0	100.0	100.0	66.7	100.0	77.2	9
@	56.0	50.0	66.0	58.7	64.7	51.3	62.2	53.3	450
V	82.8	77.1	50.0	27.7	74.3	68.6	72.7	63.6	88
eI	89.8	89.8	100.0	100.0	91.8	91.8	93.9	93.9	147
aI	86.3	86.3	94.1	90.2	94.1	94.1	91.5	89.2	153
oI	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	3
aU	87.5	87.5	87.5	87.5	93.8	87.5	89.6	87.5	48
@U	71.4	66.1	82.1	80.4	76.8	76.8	76.8	74.4	168
I@	94.1	94.1	100.0	94.1	82.4	82.4	92.2	90.2	51
e@	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	6
<at>	19.0	19.0	38.1	38.1	42.9	42.9	33.3	33.3	63
<oh>	50.0	50.0	33.3	33.3	50.0	50.0	44.4	44.4	18
<of>	45.5	45.5	27.3	27.3	45.5	45.5	39.4	39.4	33
<or>	50.0	50.0	50.0	50.0	0.0	0.0	33.3	33.3	6
Overall	75.5	68.8	82.0	76.8	80.8	76.0	79.4	73.8	6873

Table 2 Individual phoneme correct/accuracy for each and all speakers

3.1.2 Manner of articulation

There is no significant difference in the recognition performance between vowels and consonants, with vowel correctness 79.5% (n=2607) and consonants 80.6% (n=4146). However, consonants are more than twice as likely to be inserted as vowels; 267 insertions compared with 119, making the accuracy for the consonants slightly lower; consonants 74.2%, vowels 75.0%.

Again the full set of results for all speakers can be found in Appendix C (tables C5–8); only the information relevant to the current discussion will be presented here. The results analysed in terms of phoneme correctness/accuracy by manner of articulation are therefore shown in Figure 2. The overall manner class accuracy was 87.1%.

MANNER			PLACE		
Plosive	p b t d k g	1521	Labial	p b m f v T D w	879
Affricate	tS dZ	60	Alveolar	t d n s z l r	2565
Strong fricative	s z S	672	Palatal–alveolar	S tS dZ j	195
Weak fricative	f v T D h	471	Velar	k g N h	507
Liquid/Glide	l r w j	708	Front	i I E {	1032
Nasal	n m N	714	Central	V @ 3	547
Vowel	i I E { A Q O u U V @ 3 eI aI oI aU @U I@ e@	2607	Back	A O Q U u	452
			Fronting	aI eI oI	303
			Centring	I@ e@	57
			Backing	aU @U	216

Figure 1 Key of Manner and place class membership

Liquids/glides and strong fricatives were recognised most correctly and accurately for all speakers. Nasals were quite clearly the worst, especially for Speaker 1, though the accuracy of weak fricatives was also poor because of the high number of insertions. Both of these classes may be acoustically weak, and /v/ especially is easily missed, which might explain their poor performance. It is not surprising that strong fricatives should be well modelled, as they are generally acoustically prominent (compared to weak fricatives, especially). More unexpected was the good performance of liquids and glides which are often thought to be problematic for systems with limited ability to model temporal dynamics. The explanation for this may be provided by the variable frame rate analysis which is used [4]; areas which are acoustically stable are compressed into a smaller number of frames/states, while those that vary rapidly, such as /r/, /j/ and /w/ are modelled using comparatively more states, giving the improved time resolution needed to identify these sounds.

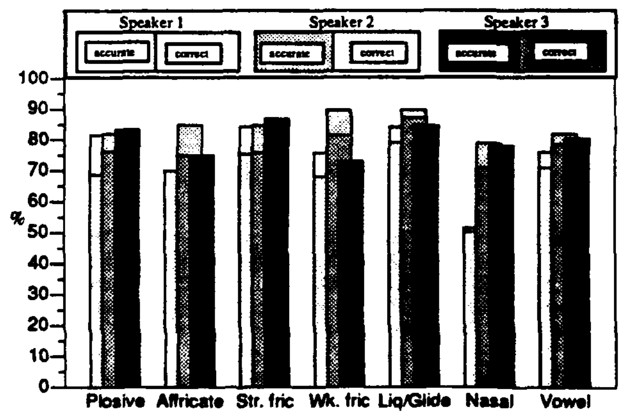

Figure 2 Graph of manner class correct/accuracy

3.1.3 Place of articulation

Figure 3 gives the analysis of the results grouped by place of articulation (and see Appendix C, tables C9–12). The overall place class accuracy was 84.4%. From the graph it can be seen that diphthongs which move towards a front position are most accurately recognised; while among the consonants, palatal–alveolars are the best recognised. Perhaps not surprisingly, central vowels were poorly dealt with. The /@/ vowel represents a large proportion (over 80%) of the central vowels and as this vowel is unstressed and notoriously variable, it is not surprising it is rather loosely modelled, and is not only easily confusable, but frequently inserted too. Labial consonants are only moderately well modelled, perhaps because most of the weak fricatives are in this group, and these are often acoustically indistinct. These results are strikingly consistent across speakers.

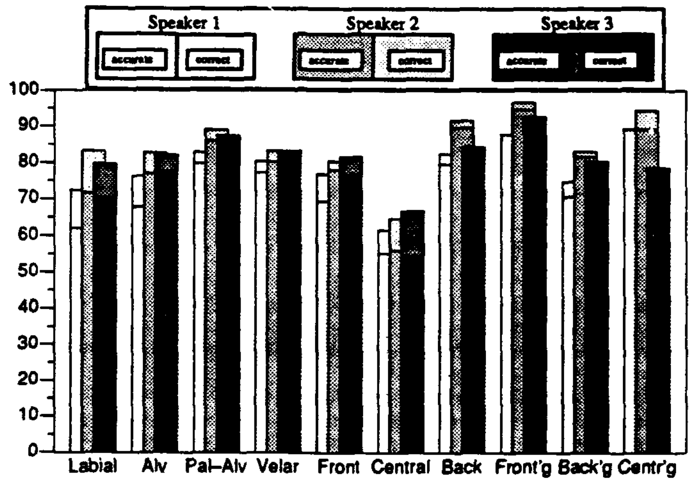

Figure 3 Graph of place class correct/accuracy

3.2 Substitutions

3.2.1 Individual phonemes

The substitution rates for individual phonemes for each and all speakers is presented in Table 3. The function words quite clearly were much substituted, and of the phonemes /N/, /V/ and/ @U/ are most likely to be substituted. /S/, /A/ and /r/ are least confused (ignoring those phonemes mentioned earlier that occur only a few times).

When the system misrecognises one phoneme as another it is important to be able to explain why this has happened. If the two phonemes involved differ minimally, in one phonetic feature (/p/ and /b/, for instance) then it may be difficult to improve either model to separate them. If, however, larger differences are involved, there may be more scope for better modelling. In order to investigate what proportion of the substitution errors were phonetically predictable, phoneme confusion matrices were constructed, and these can be found in Appendix C (tables C13–16). In general confusions are with phonetically similar sounds, though there are some exceptions to this, which are difficult to explain, even when there appears to be some pattern to them. For example, all 4 of Speaker 1's /v/–/@/ confusions occurred in the word "seven", but there were as many occasions when the /v/ in this word was correctly recognised, so it is not possible to make any generalisations about the cause of this error.

	Spkr 1	Spkr 2	Spkr 3	All	Total		Spkr 1	Spkr 2	Spkr 3	All	Total
s	11.8	6.6	7.4	8.6	408	i	12.5	13.6	4.2	11.2	224
z	10.5	21.1	7.0	12.9	171	I	18.8	21.4	15.6	18.3	394
S	3.2	0.0	0.0	1.1	93	E	18.1	9.7	4.8	10.8	249
f	5.1	5.1	12.8	7.7	234	{	7.8	10.5	12.3	10.3	165
v	23.3	10.0	6.7	13.3	90	A	2.4	0.0	5.7	2.7	111
T	2.8	14.3	16.7	11.2	107	Q	7.7	3.3	0.0	3.6	56
D	-.-	100.0	-.-	100.0	1	O	8.1	2.7	0.0	3.6	111
h	23.1	7.7	23.1	17.9	39	U	0.0	0.0	0.0	0.0	9
tS	33.3	11.1	11.1	15.4	27	u	14.6	14.6	27.3	18.8	165
dZ	18.2	18.2	18.2	18.2	33	3	0.0	0.0	0.0	0.0	9
p	4.9	4.9	12.2	7.3	123	@	14.7	18.0	18.0	16.9	450
b	13.3	6.7	33.3	17.8	45	V	14.3	33.3	17.1	21.6	88
t	5.6	6.9	10.4	7.6	693	eI	10.2	0.0	8.2	6.1	147
d	14.6	17.1	15.9	15.9	246	aI	5.9	5.9	5.9	5.9	153
k	3.1	2.1	5.2	3.4	291	oI	0.0	0.0	0.0	0.0	3
g	17.1	17.1	9.8	14.6	123	aU	12.5	12.5	6.2	10.4	48
m	32.6	10.2	4.1	15.6	147	@U	23.2	14.3	21.4	19.6	168
n	25.1	8.8	12.3	15.4	513	I@	5.9	0.0.	17.6	7.8	51
N	27.8	22.2	33.3	27.8	54	e@	50.0	50.0	50.0	50.0	6
l	5.3	9.3	10.7	8.4	225	<at>	81.0	57.1	57.1	65.1	63
r	1.9	1.0	5.8	2.9	309	<oh>	50.0	66.7	50.0	55.6	18
w	15.9	9.1	6.8	10.6	132	<of>	54.5	72.7	36.4	54.5	33
j	0.0	14.3	14.3	9.5	42	<or>	50.0	50.0	100.0	66.7	6

Table 3 Phoneme substitutions (%) for each and all speakers

3.2.2 Manner of articulation

There is no evidence that either vowels or consonants are more subject to substitution. Figure 4 shows the rate of substitutions for manner of articulation.

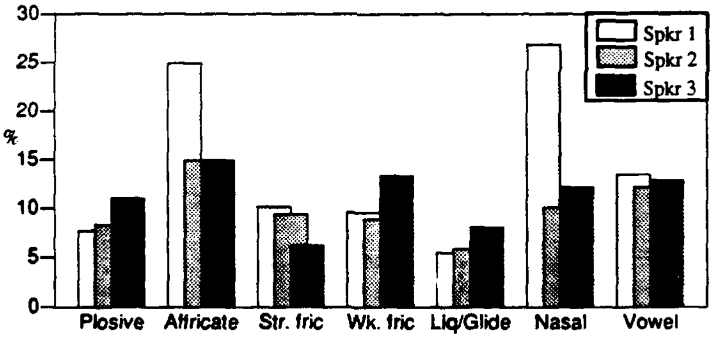

Figure 4 Graph of manner class substitutions

Confusions with phonemes from the same class would be more explicable than those with a different one, although there is a hierarchy of class similarities. For instance, plosives are more like affricates than vowels; nasals are more like vowels than they are strong fricatives. And in general this is what we find in the *ARM* results. Consonants are recognised as consonants 93%, and vowels as vowels nearly 90% of the time. The results of the finer manner class analysis of confusions for all speakers are show in Table 4 (those for individual speakers can be found, as usual in Appendix C, tables C17–20). This matrix shows how often phonemes from one class were recognised as phonemes from other classes. The matrix diagonal shows within-class recognitions.

Nasals were the most confused, though most of the confusions are predictable; nasals share stop-like characteristics with plosives, and a vowel-like structure with liquids and vowels. It is interesting that almost all (95%) of the nasal/plosive confusions were for Speaker 1, where /n/ was mostly misrecognised as /b/ or /d/.

Plosives were misrecognised most often as vowels. Nearly half of these unexpected confusions are with central vowels, indicating that /@/ is a major culprit in misrecognition (as well as being misrecognised itself). In general plosives are the most often substituted class.

		%Recognised							
		Plo	Aff	SF	WF	L/G	Nas	Vow	Total
	Plosive	87.4	0.3	0.8	0.7	0.6	0.3	1.0	137
S	Affricate	3.3	81.7	6.7	.	.	1.7	1.7	11
p	Str Fric	1.5	0.4	90.8	0.3	.	0.7	0.3	574
o	Wk Fric	5.3	0.2	0.4	80.0	.	.	1.3	362
k	Liq/Glide	0.7	0.1	.	0.6	87.3	0.4	3.7	611
e	Nasal	2.2	.	.	0.3	1.8	77.2	3.8	497
n	Vowel	0.6	.	0.1	0.2	0.6	0.5	89.9	335

Table 4 Confusion matrix for manner of articulation – all speakers

The rest of the matrix is very much as one would expect. In general in–class recognition is good. Affricates are confused with plosives and strong fricatives with which they share many features. Weak fricatives are also confused with plosives, particular confusion being /f/ with /p/, and as these share place of articulation, being broadly speaking labial, this is not unexpected.

3.2.3 Place of articulation

The substitution rates for place of articulation are shown in Figure 5. As might be expected, central vowels are the weakest; they are confused with a wide range of different classes, and are the most widely substituted class too. Backing, and for Speaker 3, centring diphthongs are also frequently confused. It can be seen from the place confusion matrix in Table 5 that much of the poor recognition of labials is likely to be due to them being confused with each other, with alveolars being the most likely substitute.

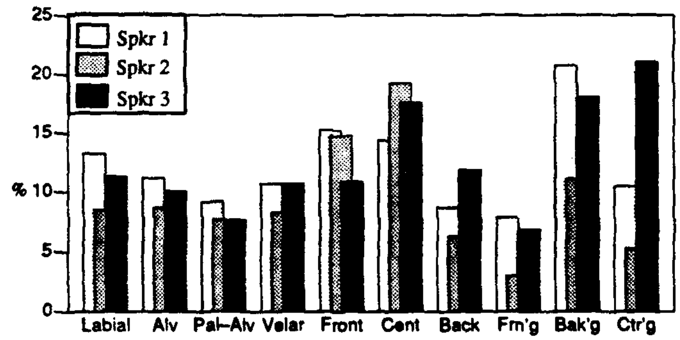

Figure 5 Graph of place class substitutions

3.2.4 Contextual Effects.

Substitution errors can sometimes be explained by the normal co–articulatory processes. Examples such as /n/ recognised as /N/ in "machine gun", /m/ as /n/ before an alveolar in "platforms", /g/ as /d/ and /d/ as /p/ in the sequence "target grid ref", /s/ as /z/ in voiced environment "zero seven", and the sequence /st/ recognised as /zd/ in the voiced environment "fuel station" are not hard to find. A more detailed description of these errors is contained in [1]. These examples account for 10% of substitutions.

In addition, as has already been mentioned above, some substitution errors are due to the quite legitimate variations which occur in fluent speech, and these nearly always involve minimal difference between target and recognised phoneme, such as place of articulation or voicing. The alternation of /i/ with /I/ in final unstressed syllables, such as in "facility", and "twenty", and /@/ with practically any unstressed vowel is well known, and was the source of on average 15% of the substitution errors. Such errors may serve to bear out the hypothesis that a major part of the substitution errors made by the system have a phonetic explanation.

		%Recognised										
		Lab	Alv	P–A	Vel	Frn	Bck	Cen	F'g	B'g	C'g	Total
Spoken	Labial	86.1	2.6	0.1	0.5	.	0.2	1.0	0.1	0.1	.	879
	Alveolar	1.9	85.4	0.3	0.2	0.3	0.2	1.2	0.2	0.1	.	2565
	Pal–Alv	.	1.0	91.3	.	1.5	.	0.5	0.5	.	.	195
	Velar	1.6	4.3	0.6	84.4	0.4	.	0.2	0.6	.	0.2	5.7
	Front	0.1	0.8	0.4	.	85.5	1.5	2.9	1.0	0.6	0.5	1032
	Back	0.6	1.1	.	.	2.7	86.5	1.1	0.2	1.3	.	547
	Central	2.2	2.0	0.2	0.2	5.7	1.5	67.1	0.5	1.1	0.4	452
	Fronting	.	.	.	.	3.0	0.7	1.0	93.1	0.7	0.3	303
	Backing	.	0.5	.	.	2.8	0.9	7.9	2.8	79.6	0.5	216
	Centring	.	.	.	.	8.8	.	1.8	.	.	89.5	57

Table 5 Confusion matrix for place of articulation – all speakers

3.3 Deletions

3.3.1 Individual phonemes

Deletions account for 42% of the recognition errors, so it would be useful to find out why they occur. Table 6 shows the deletion rates for individual phonemes.

Among the consonants /v/ scores poorly, as does /b/. We have already discussed the possible reasons for the poor performance of /v/, and of weak fricatives and nasals in general, but it is not so clear why a sound such as /b/ should be missed, but since this is consistent across speakers, it is possible that the models are defective in some way. There also appears to be a problem with /m/ specific to Speaker 1; 28.6% of this speaker's /m/s were deleted, as compared to 4.1% for both Speakers 2 and 3. There does not seem to be any particular pattern to these deletions, and there is at the moment no explanation for them, except that the models may be unreliable.

3.3.2 Manner of articulation

The deletions according to manner of articulation are shown in Figure 6. There is no real patterning to manner class deletions, although strong fricatives appear more robust for all speakers than other classes. The large percentage of nasal deletions for Speaker 1 is due in part to the predominance of /m/ deletions already mentioned, but this speaker also has nearly twice as many /n/ deletions as the others.

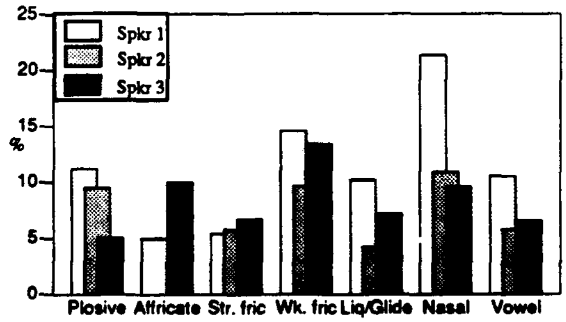

Figure 6 Graph of manner class deletions

	Spkr 1	Spkr 2	Spkr 3	All	Total		Spkr 1	Spkr 2	Spkr 3	All	Total
s	4.4	6.6	5.8	5.6	408	i	3.4	2.3	2.1	2.7	224
z	3.5	7.0	10.5	7.0	171	I	13.7	7.7	8.1	9.6	394
S	12.9	0.0	3.2	5.4	93	E	7.2	6.0	9.6	7.6	249
f	5.1	35.1	3.8	4.7	234	{	2.0	0.0	1.7	1.2	165
v	40.0	6.7	43.3	30.0	90	A	0.0	0.0	0.0	0.0	111
T	13.9	20.0	8.3	14.0	107	Q	15.4	0.0	15.4	7.1	56
D	-.-	0.0	-.-	0.0	1	O	10.8	2.7	0.0	4.5	111
h	15.4	15.4	15.4	15.4	39	U	0.0	33.3	33.3	22.2	9
tS	0.0	0.0	0.0	0.0	27	u	12.7	1.8	3.6	6.0	165
dZ	9.1	0.0	18.2	9.1	33	3	0.0	0.0	0.0	0.0	9
p	2.4	4.9	4.9	4.1	123	@	29.3	16.0	17.3	20.9	450
b	26.7	33.3	13.3	24.2	45	V	2.9	16.7	6.6	8.0	88
t	9.1	7.8	2.6	6.5	693	eI	0.0	0.0	0.0	0.0	147
d	25.6	19.5	13.4	19.5	246	aI	7.8	0.0	0.0	2.6	153
k	7.2	5.2	4.1	5.5	291	oI	0.0	0.0	0.0	0.0	3
g	7.3	4.9	2.4	4.9	123	aU	0.0	0.0	0.0	4.1	48
m	28.6	4.1	4.1	12.2	147	@U	5.4	3.6	1.8	73.6	168
n	19.9	11.1	10.5	13.8	513	I@	0.0	0.0	0.0	0.0	51
N	16.6	27.8	16.7	20.3	54	e@	0.0	0.0	0.0	0.0	6
l	16.0	4.0	13.3	11.1	225	<at>	0.0	4.8	0.0	1.6	63
r	9.7	2.9	5.8	6.2	309	<oh>	0.0	0.0	0.0	0.0	18
w	4.5	4.5	2.3	3.8	132	<of>	0.0	0.0	18.2	6.1	33
j	0.0	14.3	0.0	4.8	42	<or>	0.0	0.0	0.0	0.0	6

Table 6 Phoneme deletions (%) for each and all speakers

3.3.3 Place of articulation

Figure 7 shows the deletions analysed by place of articulation. By far the most deleted class is that of the central vowels, and this is mainly due to the phoneme /@/, which accounts for 93% of all central vowel deletions. The reasons for this are often contextual, as is discussed in 3.3.4 below. Diphthongs are not often deleted, and this could be due to the fact that they are relatively long, and usually have quite a clear structure.

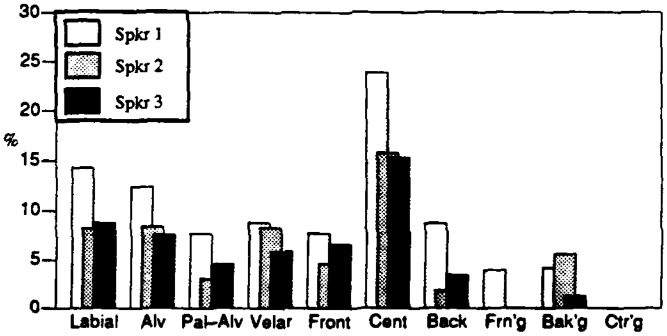

Figure 7 Graph of place class deletions

3.3.4 Contextual effects

A scored deletion is often the result of the system labelling two phonemes as one. The most typical examples of this occur when the same sound occurs at the end of one word and the beginning of the next, as in "five five" or "six six" (five is pronounced "fife" to help avoid confusion with "nine", which is pronounced "niner"). When pronounced in fluent speech the phonemes tend to run into each other, and the system recognises only one, so the above examples will be recognised as /faIfaIf/ and /sIksIks/. This is another example of how the system is penalised for an error which is due to the normal phonological processes of fluent speech. A different error of this type may be attributed to the fact that we are working with a very limited, and rather specialised vocabulary. Part of the second /n/ in "niner" is often labelled as part of the /aI/. This may be due to the fact that "niner" occurs frequently in the database, so it will have a significant influence on the (aI:n_n) triphone model (/aI/ with /n/ as its left and right context).

Many of the deletion errors are caused by genuine elisions by the speaker. For example, the unstressed /@/ vowel is often elided, particularly in unstressed syllables before a nasal or liquid. The speaker-specific dictionaries account for a number of such cases, for example in "seven" (/sEvn/) and "hidden" (/hIdn/), but in the present analysis if an /@/ appears in the dictionary transcription it will be scored as a deletion if it isn't recognised, even if in reality it wasn't produced. Words like "correction" are transcribed /k@rEkS@n/, but either (or both) of these schwas may be deleted in fast speech – /krEkSn/. It is probably for this reason that /@/ is the most deleted phoneme and is twice as likely to be deleted as any other vowel. Speaker 1 has almost twice as many /@/ deletions as the other two speakers (44 as opposed to 24 for Speaker 2 and 26 for Speaker 3), but 48% of these deletions may be attributed to legitimate variation in the way in which some words are pronounced. A slightly larger proportion of Speaker 2's schwa deletions can be so explained (58%), but less for Speaker 3 (38%).

Another case where deletion is predictable is in word-final stops, which are frequently omitted, particularly in fast speech before a word initial stop (e.g. "target category" is realised as /tAgI k{t@gri/). There are many of this type of error, and they are analysed in more detail in [1].

Deletions of /h/ are yet another example of phonologically predictable errors. This phoneme can be very variable, as it tends to take on the spectral structure of the following vowel, and is often indistinct from it. In addition one third of /h/ deletions happened after a voiceless plosive e.g. "stroke hotel" which was recognised as /str@Uk @UtEl/, where it is likely that the /h/ has been merged with the aspiration of the /k/, causing it to be missed by the system. On average 15% of all deletions can be explained by phonological effects.

3.4 Insertions

3.4.1 Individual phonemes

Insertions occur when the system has put in an extra phoneme label, and the numbers of insertions for each and all speakers is presented in Table 7. Some of the highly inserted phonemes are speaker specific – /d/ for Speaker 1 is an interesting example for which there is no explanation; others are common to all speakers, and here /@/ is the clearest example. As has already been mentioned the speaker-specific dictionaries account for a number of predictable cases of elision of this phoneme, but there are occasions when the speaker does pronounce /@/, for instance produces not a syllabic /n/, but /@n/. On average 45% of /@/ insertions could be accounted for in this way.

As an interesting aside; of the /l/ insertions, all six of those in Speaker 1's case were following the phoneme /O/. This contrasts with four out of seven, for Speaker 3 and one out of four for Speaker 2. This is interesting as for many speakers the so-called "dark /l/" resembles an /O/ vowel in spectral structure (but with a slightly lower intensity), and it could be that an off-glide of /O/ would be confused with /l/.

The phoneme /h/ was inserted only twice, in Speaker 3's data, and both after voiceless plosive (e.g. "time" recognised as /thaIm/), and we can hypothesise that the aspiration of the /t/ was what caused the insertion of /h/.

	Spkr 1	Spkr 2	Spkr 3	Total		Spkr 1	Spkr 2	Spkr 3	Total
s	11	8	4	23	i	6	3	0	9
z	9	12	3	24	I	16	1	15	32
S	0	0	0	0	E	2	1	4	7
f	3	5	2	10	{	1	3	0	4
v	5	4	4	13	A	0	1	1	2
T	4	4	2	10	Q	0	0	0	0
D	0	0	0	0	O	1	0	0	1
h	0	0	2	2	U	0	1	0	1
tS	0	0	0	0	u	3	1	1	5
dZ	0	2	0	2	3	1	0	0	1
p	14	12	10	36	@	9	11	20	40
b	2	0	4	6	V	2	4	2	8
t	8	5	10	23	eI	0	0	0	0
d	33	9	9	51	aI	0	2	0	2
k	5	4	1	10	oI	0	0	0	0
g	0	0	0	0	aU	0	0	1	1
m	2	7	3	12	@U	3	1	0	4
n	1	11	2	14	I@	0	1	0	1
N	0	1	0	1	e@	0	0	0	0
l	6	4	7	17	<at>	0	0	0	0
r	4	0	3	7	<oh>	0	0	0	0
w	0	2	2	4	<of>	0	0	0	0
j	2	0	0	2	<or>	0	0	0	0

Table 7 Phoneme insertions for each and all speakers

3.4.2 Manner of articulation

We have already mentioned that consonants are more than twice as likely to get inserted as vowels. The comparatively high level of consonant insertion was common to Speaker 2 (90 consonants compared with 30 vowels) and Speaker 1 (109 consonants and 44 vowels), but not so conspicuous in Speaker 3's results. The distribution of insertions with respect to where they occur is interesting. As many as 65% of consonant insertions were between words, perhaps being confused with breath noise or lip smacks; while 69% of vowels insertions were within words. Figure 8 shows the distribution of the rather finer manner class insertions.

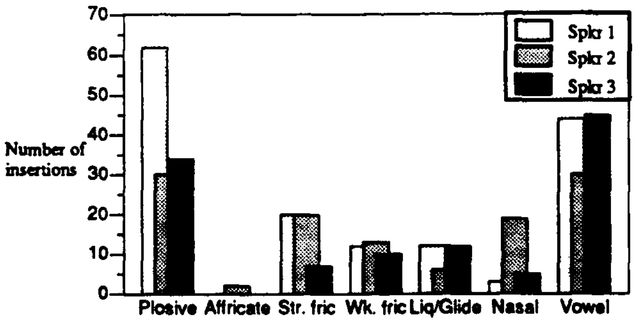

Figure 8 Graph of manner class insertions

Plosives are most frequently inserted, and 83% of these were between words, possibly bearing out the hypothesis that although there are explicit models for breath noise, lip smacks and other glitches, these sounds are nevertheless being recognised as plosives. The only class of consonants that are more often inserted (77%) within words than between them are the strong fricatives, and this may be for the same reason as vowel insertion, for which see 3.4.4 below.

3.4.3 Place of articulation

The distribution of insertions according to place of articulation is given in Figure 9. The alveolars were most likely to be inserted (though the prevalence of /d/ insertions may account for this). Diphthongs were very rarely inserted, as were the velars and palatal–alveolars, though there are too few of the latter to allow any conclusion to be drawn.

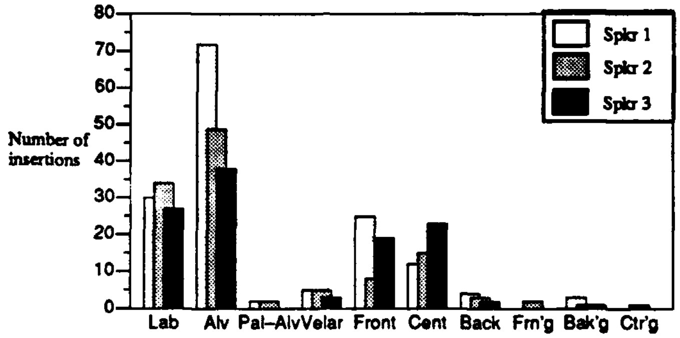

Figure 9 Graph of place class insertions

3.4.4 Contextual effects

A summary of the contexts of insertions can be found in [1]. It often happens that a long phoneme has been recognised as two separate phonemes. Sometimes these phonemes will be identical, as when /@U/ (in "zer<u>o</u>", for instance) is transcribed as /@U @U/; in other cases the insertion is phonetically related – "man<u>y</u>" is recognised as /mEnil/; or diphthongs may be recognised as two vowels, so "<u>e</u>ight" gets recognised as /elit/. Off glides from vowels are often recognised as vowel+/@/, e.g. /O/ in "f<u>our</u>" as /O@/, and /@U/ in "zer<u>o</u>" as /@U @/. Examples such as repetition of identical phoneme labels, split diphthongs, and offglide schwa account for 80% of the vowel insertions (26%, 21% and 33% respectively).

With the consonants the reasons for insertions are not so clear. Some of the insertions, like the vowels, are due to two identical labels being assigned to one phoneme (13%); others (9%) are phonetically related, as when /s/ following a voiced sound (and usually word initial) is transcribed as /z s/. For example, "four six" was recognised /fO z sIks/.

In addition, around 4% of the total consonant insertions are due to the speaker's insertion of certain sounds (mainly glides) as linkers to ease the transition between sounds. Examples of this are insertion of /r/ in "4/8" /fOreItTs/, and between "niner" and "oh". A linking /w/ is inserted in 2/8 – /tuweitTs/, and between "tango" and "eight" /t{Ng@UweIt/; and /j/ in "virtually unusable" /v3tS@lI j @n.../. The numbers of such insertions are small in our test data, but this probably reflects the limited occasions when such links could occur. For instance, /w/ was inserted only twice each by Speaker 2 and Speaker 3, and not at all by Speaker 1, but on each of those occasions it could be classed as a linking sound. Similarly, /j/ was only inserted twice by Speaker 1 and not at all by the others, and one of those insertions was a linking one. Three out of four of Speaker 1's, and two out of three of Speaker 3's /r/ insertions were linking cases. The fact that the system inserts the appropriate labels in these cases may indicate that we need to model triphones across word boundaries, rather than just word–internally as we do at the moment.

4. Discussion and conclusions

There are many interesting observations to be made from this data. What has been presented here has been an attempt to pull these together and point out general trends which might indicate what the phoneme models are doing right, as well as what they are doing wrong.

From this short discussion there have emerged two types of error: those which are genuine misrecognitions, and those which are due to the normal co-articulatory effects in fluent speech, and are thus to be expected. As far as the former are concerned, phonological effects appear to be involved in around 30% of such errors.

The vast majority of genuine errors are not unexpected, involving as they do, confusions with rather similar phonemes, or deletions of acoustically weak segments. Weak sounds such as nasals or weak fricatives predictably cause problems, as does the neutral /@/. Equally, strong and long sounds such as strong fricatives and diphthongs are well handled. The surprisingly good recognition of liquids and glides may provide an independent vindication of the use of variable frame rate analysis. A large number of the insertions and deletions could probably be prevented if our duration modelling were more sophisticated.

Although the majority of the errors appear to have a phonetic basis, there are cases where the errors are as yet inexplicable from a phonetic point of view – the unusually large number of /d/ insertions in Speaker 1's data, and the poor recognition of the same speaker's /m/, Speaker 3's /b/ and Speaker 2's /V/ for example. A small number of phonemes (and Speaker 2's /V/ may belong to this group) may simply not occur frequently enough for a reliable indication of performance to be made. Where there isn't a phonetic explanation of an error, it would be interesting to find out if the system's own measure of its goodness of match is consistent with our judgement of its performance.

It is important to remember that this study was based on a system which used no dictionary, although the triphones are forced to match at the edges. When lexical and syntactic constraints are available, as they are when the system is run in its usual mode, as a word recogniser, then many of the problems discussed above no longer occur. However, a general improvement in the sub-word level modelling would provide a sound basis for better word recognition and this study has enabled us to pinpoint a few areas where our models might be improved, and may indicate that we need to give some consideration to phonological effects across word boundaries. The level of performance depends ultimately depends on the task and vocabulary, and a next step might be to assess the extent to which the somewhat specialised vocabulary of the *ARM* task has influenced these results, by looking at other tasks, and bigger vocabularies, as well as at a wider range of speakers.

References

[1] J D Crowe, "Contextual influences on phoneme recognition in the *ARMADA* speech recognition system", RSRE SP4 Research Note No. 129, 1990.

[2] A J Fourcin, G Harland, W Barry,and V Hazan, "Speech Input and Output Assessment. Multilingual Methods and Standards", Ellis Horwood Ltd., Chichester, 1989.

[3] A C Gimson, An introduction to the pronunciation of English. (Second Edition) Edward Arnold (Publishers) Ltd. 1970.

[4] S M Peeling and K M Ponting, "Further experiments in variable frame rate analysis for speech recognition", RSRE Memorandum No. 4336, 1990.

[5] K M Ponting and M J Russell, "The ARM project: automatic recognition of spoken airborne reconnaissance reports", Military and Government Speech Tech '89, Arlington Va., 1989.

[6] M J Russell, K M Ponting, S R Browning, S Downey and P Howell, "Triphone clustering in the *ARM* system", RSRE Memorandum No. 4357, 1990.

[7] M J Russell, K M Ponting, S M Peeling, S R Browning, J S Bridle, R K Moore, I Galiano and P Howell, "The *ARM* Continuous Speech Recognition System", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Albuquerque, New Mexico, 1990.

[8] J Wells et al., "Specification of the SAM Phonetic Alphabet (SAM-PA)", included in P Winski, W J Barry and A Fourcin (Eds.) "Support available from the SAM Project for other ESPRIT speech and language work", the SAM Project, University College London, 1990.

INTENTIONALLY BLANK

Appendix A. Phoneme frequency

Rank	Frequency [3]	Phoneme	Example	ARM frequency	ARM rank
1	10.74	@	alpha	6.67	3
2	8.33	I	civil	5.80	5
3	7.58	n	new	7.60	2
4	6.42	t	target	10.27	1
5	5.14	d	damaged	3.64	10
6	4.81	s	six	6.04	4
7	3.66	l	lima	3.33	12
8	3.56	D	then	0.01	42
9	3.51	r	rail	4.58	6
10	3.22	m	map	2.18	=18
11	3.09	k	correction	4.31	7
12	2.97	E	enemy	3.67	9
13	2.81	w	well	1.96	20
14	2.46	z	comprising	2.53	13
15	2.00	v	over	1.33	27
16	1.97	b	about	0.67	33
17	1.83	aI	dimension	2.27	17
18	1.79	f	fife	3.46	11
19	1.78	p	papa	1.82	=21
20	1.75	V	up	1.30	28
21	1.71	eI	containing	2.18	=18
22	1.65	i	beacon	3.91	8
23	1.51	@U	close	2.49	14
24	1.46	h	hotel	0.58	=35
25	1.45	{	damaged	2.44	=15
26	1.37	Q	approx	0.83	29
27	1.24	O	four	1.64	=23
28	1.15	N	bearing	0.80	30
29	1.13	u	two	2.44	=15
30	1.05	g	grass	1.82	=21
31	0.96	S	ambush	1.38	26
32	0.88	j	yards	0.62	34
33	0.86	U	woods	0.13	=38
34	0.79	A	charlie	1.64	=23
35	0.61	aU	south	0.71	32
36	0.60	dZ	damaged	0.49	36
37	0.52	3	heard	0.13	=38
38	0.41	tS	charlie	0.40	37
39	0.37	T	three	1.58	25
40	0.34	e@	air	0.08	40
41	0.21	I@	clear	0.76	31
42	0.14	oI	destroyed	0.04	41
43	0.10	Z	pleasure	0.00	=43
44	0.06	U@	poor	0.00	=43

INTENTIONALLY BLANK

Appendix B. The ARM vocabulary

about	@baUt	above	@bVv
access	{ksEs	ack–ack	{k{k
acquisition	{kwIzIS@n	action	{kS@n
active	{kUv	activities	{kUvItiz
activity	{kUvIti	aerial	e@rI@l
aerials	e@rI@lz	AEW	eIidVb@lju
aircraft	e@krAft	airfield	e@fi@ld
air–raid	e@reId	airstrip	e@strIp
alpha	{lf@	ambush	{mbUS
ammo	{m@U	ammunition	{mjunIS@n
and	{nd	antenna	{ntEn@
antennae	{ntEnaI	anti–aircraft	{ntie@krAft
anti–armour	{ntiAm@	anti–tank	{ntit{nk
approx	@prQks	approximately	@prQksIm@tli
APVs	eIpiviz	armour	Am@
armoured	Am@d	arms	Amz
array	@reI	artillery	AtIl@ri
as	{z	assembled	@sEmbld
assembly	@sEmbli	associated	@s@UsieIt@d
badger	b{dZ@	be	bi
beacon	bik@n	bear	be@
bearing	be@rIN	being	biIN
below	bIl@U	blocked	blQkt
bomber	bQm@	bowsers	baUz@z
bravo	brAv@U	bridge	brIdZ
brow	braU	C2	situ
camouflaged	k{m@flAZd	camp	k{mp
canal	k@n{l	cantilever	k{ntIliv@
carriage	k{rIdZ	carriages	k{rIdZIz
capability	keIp@bIlIti	capacity	k@p{sIti
casevac	k{z@v{k	cat	k{t
category	k{t@gri	centre	sEnt@
charlie	tSAli	circular	s3kjul@
civil	sIv@l	civilian	s@vIlj@n
clear	klI@	close	kl@Us
comms	kQmz	collection	k@lEkS@n
column	kQl@m	communications	k@mjunIkeIS@nz
complete	k@mplit	comprehensive	kQmprIhEnsIv
comprising	k@mpraIzIN	concealed	k@nsild
concrete	kQnkrit	conical	kQnIk@l
considerable	k@nsIdr@bl	consisting	k@nsIstIN
construction	k@nstrVkS@n	containing	k@nteInIN
convoy	kQnvoI	co–ordinates	k@UOdIn@ts
co–ords	k@UOdz	correction	k@rEkS@n
covert	k@Uv3t	crossing	krQsIN
damaged	d{mIdZd	dash	d{S
data	deIt@	deck	dEk
defence	dIfEns	defences	dIfEnsIz
defended	dIfEndId	degrees	d@griz
delta	dElt@	depot	dEp@U
diameter	daI{mIt@	difficult	dIfIk@lt
dimension	daImEnS@n	dipole	daIp@Ul
dipoles	daIp@Ulz	direct	dIrEkt
direction	dIrEkS@n	dish	dIS
dispersal	dIsp3s@l	dispersals	dIsp3s@lz
destroyed	dIstroId	dug	dVg
dump	dVmp	each	itS
easy	izi	east	ist
echo	Ek@U	eight	eIt
eighteen	eItin	eighty	eIti
eleven	IlEv@n	emplacement	EmpleIsm@nt
end	End	enemy	En@mi

Appendix B. The ARM vocabulary

engineering	EndZInI@rIN	erect	IrEkt
ESM	iEsEm	estimated	EstImeItId
evidence	EvId@ns	evident	EvId@nt
EW	idVblju	facility	f@sIlIti
feet	fit	a_few	@fju
field	fi@ld	fifteen	fIftin
fifty	fIfti	*fighter*	faIt@
fire	faI@	firing	faI@rIN
fishbed	fISbEd	five	faIf
fixed	fIkst	flanker	fl{nk@
flogger	flQg@	forger	fOdZ@
fortified	fOtIfaId	forty	fOti
four	fO	fourteen	fOtin
foxbat	fQksb{t	foxtrot	fQkstrQt
freight	*freIt*	frenzied	frEnzid
fuel	fju@l	fulcrum	fUlkr@m
GMT	dZiEmti	golf	gQlf
going	g@UIN	goods	gUdz
grass	grAs	grid	grId
ground	graUnd	guidance	gaId@ns
gun	gVn	guns	gVnz
hangar	h{N@	hardened	hAd@nd
having	h{vIN	havok	h{v@k
heading	hEdIN	heavy	hEvi
height	haIt	helicopters	hElIkQpt@z
helos	hEl@Uz	hidden	hIdn
hind	haInd	hip	hIp
hocum	h@Uk@m	holding	h@UldIN
horizontal	hQrIzQnt@l	horn	hOn
hospital	hQspIt@l	hotel	h@UtEl
hour	aU@	howitzer	haUwItz@
hundred	hVndrEd	hurried	hVrid
hyphen	haIf@n	including	InkludIN
incomplete	Ink@mplit	incorporating	InkOp@reItIN
india	Indj@	infantry	Inf@ntri
inoperative	InQpr@tIv	*installation*	Inst@leIS@n
intact	Int{kt	intelligence	IntElIdZ@ns
joint	dZoInt	juliet	dZuliEt
junction	dZVnkS@n	kilo	kil@U
kilometres	kIlQm@t@z	kilometres/hour	kIlQm@t@zp3aU@
knots	nQts	lanes	leInz
launch	lOntS	launcher	lOntS@
launchers	lOntS@z	length	lENT
less	lEs	level	lEv@l
lift	lIft	light	laIt
like	laIk	lima	lim@
limited	lIm@tId	lines	laInz
little	lItl	loading	l@UdIN
located	l@UkeItId	location	l@UkeIS@n
logistics	lQdZIstIks	loop	lup
lorry	lQri	machine	m@Sin
machine-gun	m@SingVn	main	meIn
maintenance	meInt@n@ns	major	meIdZ@
many	mEni	map	m{p
marshalling	mAS@lIN	mast	mAst
material	m@tI@rI@l	MCVs	Emsiviz
mechanised	mEk@naIzd	medium	midj@m
message	mEsIdZ	metres	mit@z
mike	maIk	miles	maIlz
miles/hour	maIlzp3aU@	military	mIl@tri
minor	maIn@	missile	mIsaIl
missiles	mIsaIlz	mixed	mIkst

Appendix B. The ARM vocabulary

mobile	m@Uball
more	mO
motorised	m@Ut@ralzd
movement	muvm@nt
navaid	n{veld
new	nju
nineteen	nalntin
no	n@U
north	nOT
northwest	nOTwEst
noticed	n@Utlst
number	nVmb@
numerous	njum@r@s
obstructed	QbstrVktld
on	Qn
operational	Qp@relS@n@l
out	aUt
pack	p{k
parabolic	p{r@bQllk
partly	pAtli
peak	pik
perhaps	p@h{ps
personnel	p3s@nEl
platforms	pl{tfOmz
plus	plVs
pontoon	pQntun
possibly	pQslbli
prepared	prlpe@d
proceeding	pr@sidlN
protected	pr@tEktld
radar	reldA
rail	rell
ready	rEdi
recce	rEki
reconnaissance	r@kQnls@ns
ref	rEf
re-fuelling	rifju@llN
repair	r@pe@
repeat	r@pit
rhombic	rQmblk
road	r@Ud
rockets	rQklts
rotary	r@Ut@ri
rounds	raUndz
runways	rVnwelz
SAM-7	s{msEvn
scout	skaUt
sections	sEkS@nz
semi	sEmi
sets	sEts
seventeen	sEvntin
several	sEvr@l
sidings	saldlNz
sighted	saltld
similar	slmll@
site	salt
sixteen	slkstin
size	salz
slash	sl{S
so	s@U
something	sVmTIN
southeast	saUTist
modified	mQdlfald
most	m@Ust
motorway	m@Ut@wel
much	mVtS
near	nl@
nine	naln@
ninety	nalnti
normal	nOm@l
northeast	nOTist
not	nQt
november	n@UvEmb@
a-number	@nVmb@
observed	Qbz3vd
occupied	Qkjupald
one	wVn
oscar	Qsk@
over	@Uv@
papa	p{p{
partially	pAS@li
passenger	p{s@ndZ@
per	p3
permanent	p3m@n@nt
pipeline	palplaln
platoon	pl@tun
police	p@lis
position	p@zlS@n
practically	pr{ktlkli
principal	prlnslp@l
projectiles	pr@dZEktallz
quebec	kw@bEk
radio	reldj@U
railway	rellwel
re-arming	riAmlN
receiver	r@siv@
red-cross	rEdkrQs
reference	rEfr@ns
refurbishing	rlf3blSIN
repaired	r@pe@d
report	r@pOt
river	rlv@
rocket	rQklt
romeo	r@Umi@U
roughly	rVfli
runway	rVnwel
SAM	s{m
scientific	sal@ntlflk
section	sEkS@n
self-propelled	sElfpr@pEld
serviceable	s3vls@bl
seven	sEvn
seventy	sEv@nti
siding	saldlN
sierra	siEr@
sighting	saltlN
single	slNgl
six	slks
sixty	slksti
skip	sklp
small	smOl
some	sVm
south	saUT
southwest	saUTwEst

Appendix B. The ARM vocabulary

span	sp{n	spans	sp{nz
speed	spid	SPGs	EspidZiz
spotted	spQtId	squad	skwQd
squadron	skwQdr@n	static	st{tIk
station	steIS@n	steel	sti@l
stone	st@Un	stop	stQp
storage	stOrIdZ	stores	stOz
strip	strIp	stroke	str@Uk
structural	strVktS@r@l	summit	sVmIt
supply	s@plaI	surface	s3fIs
support	s@pOt	suspension	s@spEnS@n
swing	swIN	tactical	t{ktIk@l
tango	t{Ng@U	tanker	t{Nk@
tankers	t{Nk@z	tanks	t{Nks
target	tAgIt	TARWI	tAwi
task	tAsk	taxiway	t{ksIweI
taxiways	t{ksIweIz	temporarily	tEmp@re@rIli
temporary	tEmp@ri	ten	tEn
than	D@n	thirteen	T3tin
thirty	T3ti	thousand	TaUz@nd
three	Tri	time	taIm
to	tu	total	t@Ut@l
tracked	tr{kt	tracks	tr{ks
train	treIn	trains	treInz
transmitter	trAnzmIt@	transport	trAnspOt
trees	triz	troop	trup
troops	trups	twelve	twElv
twenty	twEnti	twenty-one	twEntiwVn
twenty-three	twEntiTri	twenty-two	twEntitu
twin	twIn	two	tu
type	taIp	undamaged	Vnd{mIdZd
undefended	VndIfEndId	under	Vnd@
unidentified	VnaIdEntIfaId	unipole	junIp@Ul
unknown	Vn@Un	unloading	Vnl@UdIN
unobstructed	Vn@bstrVktId	unoccupied	VnQkjupaId
unoperational	VnQp@reIS@n@l	unrepaired	Vnr@pe@d
unserviceable	Vns3vIs@bl	unusable	Vnjuz@bl
uniform	junIfOm	up	Vp
U/S	juEs	usable	juz@bl
use	jus	vehicle	vi@kl
vehicles	vi@klz	vertical	v3tIkl
victor	vIkt@	virtually	v3tS@li
visible	vIz@bl	VSTOL	vIstQl
wagon	w{g@n	wagons	w{g@nz
water	wOt@	weapon	wEp@n
weapons	wEp@nz	well	wEl
west	wEst	whiskey	wIski
wholly	h@Uli	wing	wIN
wire	waI@	width	wIdT
with	wIT	wood	wUd
wooden	wUd@n	woods	wUdz
work	w3k	worked	w3kt
x-ray	EksreI	YAGI	jAgi
yankee	j{Nki	yard	jAd
yards	jAdz	zero	zI@r@U
zulu	zulu		

Appendix C. Results in full

Phoneme	Total	% Cor	% Sub	% Del	No. of Ins	% Acc	Phoneme	Total	Cor	Sub	Del	Ins	Acc
s	136	83.8	11.8	4.4	11	75.7	i	88	84.1	12.5	3.4	6	77.3
z	57	86.0	10.5	3.5	9	70.2	I	117	65.5	18.8	13.7	16	53.8
S	31	83.9	3.2	12.9	0	83.9	E	83	74.7	18.1	7.2	2	72.3
f	78	89.8	5.1	5.1	3	85.9	{	51	90.2	7.8	2.0	1	88.2
v	30	36.7	23.3	40.0	5	20.0	A	41	97.6	2.4	0.0	0	97.6
T	36	83.3	2.8	13.9	4	72.2	Q	13	76.9	7.7	15.4	0	76.9
D	0	-.-	-.-	-.-	-	-.-	O	37	81.1	8.1	10.8	1	78.4
h	13	61.5	23.1	15.4	0	61.5	U	3	100.0	0.0	0.0	0	100.0
tS	9	66.7	33.3	0.0	0	66.7	u	55	72.7	14.6	12.7	3	67.3
dZ	11	72.7	18.2	9.1	0	72.7	3	3	100.0	0.0	0.0	1	66.7
p	41	92.7	4.9	2.4	14	58.5	@	150	56.0	14.7	29.3	9	50.0
b	15	60.0	13.3	26.7	2	46.7	V	35	82.8	14.3	2.9	2	77.1
t	231	85.3	5.6	9.1	8	81.8	eI	49	89.8	10.2	0.0	0	89.8
d	82	59.8	14.6	25.6	33	19.5	aI	51	86.3	5.9	7.8	0	86.3
k	97	89.7	3.1	7.2	5	84.5	oI	1	100.0	0.0	0.0	0	100.0
g	41	75.6	17.1	7.3	0	75.6	aU	16	87.5	12.5	0.0	0	87.5
m	49	38.8	32.6	28.6	2	34.7	@U	56	71.4	23.2	5.4	3	66.1
n	171	55.0	25.1	19.9	1	54.4	I@	17	94.1	5.9	0.0	0	94.1
N	18	55.6	27.8	16.6	0	55.6	e@	2	50.0	50.0	0.0	0	50.0
l	75	78.7	5.3	16.0	6	70.7	<aI>	21	19.0	81.0	0.0	0	19.0
r	103	88.4	1.9	9.7	4	84.5	<oh>	6	50.0	50.0	0.0	0	50.0
w	44	79.6	15.9	4.5	0	79.6	<of>	11	45.5	54.5	0.0	0	45.5
j	14	100.0	0.0	0.0	2	85.7	<or>	2	50.0	50.0	0.0	0	50.0

Table C1. Phoneme recognition results for Speaker 1.

Phoneme	Total	% Cor	% Sub	% Del	No. of Ins	% Acc	Phoneme	Total	% Cor	% Sub	% Del	No. of Ins	% Acc
s	136	86.8	6.6	6.6	8	80.9	i	88	84.1	13.6	2.3	3	80.6
z	57	71.9	21.1	7.0	12	50.8	I	117	70.9	21.4	7.7	1	70.1
S	31	100.0	0.0	0.0	0	100.0	E	83	84.3	9.7	6.0	1	83.1
f	78	89.8	5.1	5.1	5	83.3	{	57	89.5	10.5	0.0	3	84.2
v	30	83.3	10.0	6.7	4	70.0	A	35	100.0	0.0	0.0	1	97.1
T	35	65.7	14.3	20.0	4	54.3	Q	30	96.7	3.3	0.0	0	96.7
D	1	0.0	100.0	0.0	0	0.0	O	37	94.6	2.7	2.7	0	94.6
h	13	76.9	7.7	15.4	0	76.9	U	3	66.7	0.0	33.3	1	33.3
tS	9	88.9	11.1	0.0	0	88.9	u	55	83.6	14.6	1.8	1	81.8
dZ	11	81.8	18.2	0.0	2	63.6	3	3	100.0	0.0	0.0	0	100.0
p	41	90.2	4.9	4.9	12	61.0	@	150	66.0	18.0	16.0	11	58.7
b	15	60.0	6.7	33.3	0	60.0	V	18	50.0	3.3	16.7	4	27.7
t	231	85.3	6.9	7.8	5	83.1	eI	49	100.0	0.0	0.0	0	100.0
d	82	63.4	17.1	19.5	9	52.4	aI	51	94.1	5.9	0.0	2	90.2
k	97	92.7	2.1	5.2	4	88.7	oI	1	100.0	0.0	0.0	C	100.0
g	41	78.0	17.1	4.9	0	78.0	aU	16	87.5	12.5	0.0	0	87.5
m	49	85.7	10.2	4.1	7	71.4	@U	56	82.1	14.3	3.6	1	80.4
n	171	80.1	8.8	11.1	11	73.7	I@	17	100.0	0.0	0.0	1	94.1
N	18	50.0	22.2	27.8	1	44.4	e@	2	50.0	50.0	0.0	0	50.0
l	75	86.7	9.3	4.0	4	81.3	<aI>	21	38.1	57.1	4.8	0	38.1
r	103	96.1	1.0	2.9	0	96.1	<oh>	6	33.3	66.7	0.0	0	33.3
w	44	86.4	9.1	4.5	2	81.8	<of>	11	27.3	72.7	0.0	0	27.3
j	14	71.4	14.3	14.3	0	71.4	<or>	2	50.0	50.0	0.0	0	50.0

Table C2. Phoneme recognition results for Speaker 2.

Appendix C. Results in full

Phoneme	Total	% Cor	% Sub	% Del	No. of Ins	% Acc	Phoneme	Total	% Cor	% Sub	% Del	No. of Ins	% Acc
s	136	86.8	7.4	5.8	4	83.8	i	48	93.7	4.2	2.1	0	93.7
z	57	82.5	7.0	10.5	3	77.2	I	160	76.3	15.6	8.1	15	66.9
S	31	96.8	0.0	3.2	0	96.8	E	83	85.6	4.8	9.6	4	80.7
f	78	83.4	12.8	3.8	2	80.8	{	57	86.0	12.3	1.7	0	86.0
v	30	50.0	6.7	43.3	4	36.7	A	35	94.3	5.7	0.0	1	91.4
T	36	75.0	16.7	8.3	2	69.4	Q	13	84.6	0.0	15.4	0	84.6
D	0	-.-	-.-	-.-	-	-.-	O	37	100.0	0.0	0.0	0	100.0
h	13	61.5	23.1	15.4	2	46.2	U	3	66.7	0.0	33.3	0	66.7
tS	9	88.9	11.1	0.0	0	88.9	u	55	69.1	27.3	3.6	1	67.3
dZ	11	63.6	18.2	18.2	0	63.6	3	3	100.0	0.0	0.0	0	66.7
p	41	82.9	12.2	4.9	10	58.5	@	150	64.7	18.0	17.3	20	51.3
b	15	53.4	33.3	13.3	4	26.7	V	35	74.3	17.1	8.6	2	68.6
t	231	87.0	10.4	2.6	10	82.7	eI	49	91.8	8.2	0.0	0	91.8
d	82	70.7	15.9	13.4	9	59.8	aI	51	94.1	5.9	0.0	0	94.1
k	97	90.7	5.2	4.1	1	89.7	oI	1	100.0	0.0	0.0	0	100.0
g	41	87.8	9.8	2.4	0	87.8	aU	16	93.8	6.2	0.0	1	87.5
m	49	91.8	4.1	4.1	3	85.7	@U	56	76.8	21.4	1.8	0	76.8
n	171	77.2	12.3	10.5	2	76.0	I@	17	82.4	17.6	0.0	0	82.4
N	18	50.0	33.3	16.7	0	50.0	e@	2	50.0	50.0	0.0	0	50.0
l	75	76.0	10.7	13.3	7	66.7	<ar>	21	42.9	57.1	0.0	0	42.9
r	103	88.4	5.8	5.8	3	85.4	<oh>	6	50.0	50.0	0.0	0	50.0
w	44	90.9	6.8	2.3	2	86.4	<of>	11	45.4	36.4	18.2	0	45.4
j	14	85.7	14.3	0.0	0	85.7	<or>	2	0.0	100.0	0.0	0	0.0

Table C3. Phoneme recognition performance for Speaker 3.

Phoneme	Total	% Cor	% Sub	% Del	No. of Ins	% Acc	Phoneme	Total	% Cor	% Sub	% Del	No. of Ins	% Acc
s	408	85.8	8.6	5.6	23	80.2	i	224	86.1	11.2	2.7	9	82.1
z	171	80.1	12.9	7.0	24	65.5	I	394	72.1	18.3	9.6	32	64.0
S	93	93.5	1.1	5.4	0	93.5	E	249	81.5	10.8	7.6	7	78.7
f	234	87.6	7.7	4.7	10	83.3	{	165	88.5	10.3	1.2	4	86.1
v	90	56.7	13.3	30.0	13	42.3	A	111	97.3	2.7	0.0	2	95.5
T	107	74.8	11.2	14.0	10	65.5	Q	56	89.3	3.6	7.1	0	89.3
D	1	0.0	100.0	0.0	0	0.0	O	111	91.9	3.6	4.5	1	91.0
h	39	66.7	17.9	15.4	2	61.6	U	9	77.8	0.0	22.2	1	66.7
tS	27	81.5	18.5	0.0	0	81.5	u	165	75.2	18.8	6.0	5	72.2
dZ	33	72.7	18.2	9.1	2	66.6	3	9	100.0	0.0	0.0	2	77.2
p	123	88.6	7.3	4.1	36	59.3	@	450	62.2	16.9	20.9	40	53.3
b	45	57.9	17.8	24.2	6	44.4	V	88	72.7	19.3	8.0	8	63.6
t	693	85.9	7.6	6.5	23	82.6	eI	147	93.9	6.1	0.0	0	93.9
d	246	64.6	15.9	19.5	51	43.9	aI	153	91.5	5.9	2.6	2	89.2
k	291	91.1	3.4	5.5	10	87.0	oI	3	100.0	0.0	0.0	0	100.0
g	123	80.5	14.6	4.9	0	80.5	aU	48	89.6	6.3	4.1	1	87.5
m	147	72.2	15.6	12.2	12	64.0	@U	168	76.8	19.6	3.6	4	74.4
n	513	70.8	15.4	13.8	21	66.7	I@	51	92.2	7.8	0.0	1	90.2
N	54	51.9	27.8	20.3	1	50.0	e@	6	50.0	50.0	0.0	0	50.0
l	225	80.5	8.4	11.1	17	72.9	<ar>	63	33.3	65.1	1.6	0	33.3
r	309	90.9	2.9	6.2	7	88.7	<oh>	18	44.4	55.6	0.0	0	44.4
w	132	85.6	10.6	3.8	4	82.6	<of>	33	39.4	54.5	6.1	0	39.4
j	42	85.7	9.5	4.8	2	80.9	<or>	6	33.3	66.7	0.0	0	33.3

Table C4. Phoneme recognition performance for all speakers combined.

Appendix C. Results in full

Class	% Cor	% Sub	% Del	No. of Ins	% Acc	Total
Plosive	81.1	7.7	11.2	62	65.8	507
Affricate	70.0	25.0	5.0	0	70.0	20
Str Fric	84.4	10.2	5.4	20	75.4	224
Wk Fric	75.8	9.6	14.6	12	68.2	157
Liq/Glide	84.3	5.5	10.2	12	79.2	236
Nasal	51.7	26.9	21.4	3	50.4	238
Vowel	76.0	13.5	10.5	44	71.0	868
Average	74.8	14.1	9.8	21.9	68.6	2250

Table C5. Manner results for Speaker 1.

Class	% Cor	% Sub	% Del	No. of Ins	% Acc	Total
Plosive	82.2	8.3	9.5	30	76.3	507
Affricate	85.0	15.0	0.0	2	75.0	20
Str Fric	84.8	9.4	5.8	20	75.9	224
Wk Fric	81.5	8.9	9.6	13	73.2	157
Liq/Glide	89.9	5.9	4.2	6	87.3	236
Nasal	79.0	10.1	10.9	19	71.0	238
Vowel	82.0	12.2	5.8	30	78.6	868
Average	83.5	10.0	6.5	17.1	76.7	2250

Table C6. Manner results for Speaker 2.

Class	% Cor	% Sub	% Del	No. of Ins	% Acc	Total
Plosive	83.8	11.1	5.1	34	77.1	507
Affricate	75.0	15.0	10.0	0	75.0	20
Str Fric	87.0	6.3	6.7	7	83.9	224
Wk Fric	73.2	13.4	13.4	10	66.9	157
Liq/Glide	84.7	8.1	7.2	12	79.9	236
Nasal	78.2	12.2	9.6	5	76.1	238
Vowel	80.5	12.9	6.6	45	75.3	871
Average	80.3	11.3	8.4	16.1	76.3	2253

Table C7. Manner results for Speaker 3.

Class	% Cor	% Sub	% Del	No. of Ins	% Acc	Total
Plosive	82.4	9.0	8.6	126	74.1	1521
Affricate	76.7	18.3	5.0	2	73.3	60
Str Fric	85.4	8.6	6.0	47	78.4	672
Wk Fric	76.9	10.6	12.5	35	69.4	471
Liq/Glide	86.3	6.5	7.2	30	82.1	708
Nasal	69.6	16.4	14.0	27	65.8	714
Vowel	79.5	12.9	7.6	119	75.0	2607
Average	79.5	10.6	8.7	55.1	74.0	6753

Table C8. Manner results for all speakers.

Appendix C. Results in full

Class	% Cor	% Sub	% Del	No. of Ins	% Acc	Total
Labial	72.4	13.3	14.3	30	62.1	293
Alveolar	76.4	11.2	12.4	72	68.0	855
Pal-Al	83.1	9.2	7.7	2	80.0	65
Velar	80.5	10.7	8.8	5	77.5	169
Front	77.0	15.3	7.7	25	69.6	339
Central	61.7	14.4	23.9	12	55.3	188
Back	82.6	8.7	8.7	4	79.9	149
Fronting	88.1	7.9	4.0	0	88.1	101
Backing	75.0	20.8	4.2	3	70.8	72
Centring	89.5	10.5	0.0	0	89.5	19
Average	78.6	12.2	9.1	15.3	74.1	2250

Table C9. Place results for Speaker 1.

Class	% Cor	% Sub	% Del	No. of Ins	% Acc	Total
Labial	83.3	8.5	8.2	34	71.7	293
Alveolar	82.9	8.7	8.4	49	77.2	855
Pal-Al	89.2	7.7	3.1	2	86.2	65
Velar	83.4	8.3	8.3	5	80.5	169
Front	80.6	14.8	4.6	8	78.3	345
Central	64.9	19.3	15.8	15	56.1	171
Back	91.9	6.3	1.8	3	90.0	160
Fronting	97.0	3.0	0.0	2	95.0	101
Backing	83.3	11.1	5.6	1	81.9	72
Centring	94.7	5.3	0.0	1	89.5	19
Average	85.1	9.3	5.6	12.0	80.6	2250

Table C10. Place Results for Speaker 2.

Class	% Cor	% Sub	% Del	No. of Ins	% Acc	Total
Labial	79.9	11.3	8.8	27	70.6	293
Alveolar	82.3	10.1	7.6	38	77.9	855
Pal-Al	87.7	7.7	4.6	0	87.7	65
Velar	83.4	10.7	5.9	3	81.7	169
Front	82.5	10.9	6.6	19	77.0	348
Central	67.0	17.6	15.4	23	54.8	188
Back	84.6	11.9	3.5	2	83.2	143
Fronting	93.1	6.9	0.0	0	93.1	101
Backing	80.6	18.1	1.3	1	79.2	72
Centring	78.9	21.1	0.0	0	78.9	19
Average	82.0	12.6	5.4	11.3	78.4	2253

Table C11. Place Results for Speaker 3.

Class	% Cor	% Sub	% Del	No. of Ins	% Acc	Total
Labial	78.5	11.0	10.5	91	68.1	879
Alveolar	80.5	10.0	9.5	159	74.3	2565
Pal-Al	86.7	8.2	5.1	4	84.6	195
Velar	82.4	9.9	7.7	13	79.9	507
Front	80.0	13.7	6.3	52	75.0	1032
Central	64.5	17.1	18.4	50	55.4	547
Back	86.6	8.8	4.6	9	84.5	452
Fronting	92.8	5.9	1.3	2	92.1	303
Backing	79.6	16.7	3.7	5	77.3	216
Centring	87.7	12.3	0.0	1	86.0	57
Average	73.9	11.5	6.7	38.6	77.7	6753

Table C12. Place results for all speakers.

Appendix C. Results in full

	s	S	z	Z	f	v	T	D	h	tS	dZ	p	b	t	d	k	g	m	n	N	l	r	w	j	i	I	E	{	A	Q	O	U	u	3	@	V	eI	aI	oI	aU	@U	I@	e@	U@	<ax>	<ah>	<af>	<an>	<ux>	Total
s	114		11				1			1				3																																				136
S		26									1																																							31
z	3		49											3																																				57
Z				0																																														0
f					70		2					2																																						78
v						11						2																							4												1			30
T		1					30																																											36
D								0																																										0
h									8	1					1	1																																		13
tS		2								6				1																																				9
dZ		1									8								1																															11
p					1				1			38																																						41
b												1	9																						1															15
t	3		2											197	4	1									1								1												1					231
d	1		1								1			5	49											1									2		1													82
k										1		2				87																																		97
g												1			1	4	31					1																												41
m						2							5					19	7		1	1																												49
n													7		7			12	94	2	[illegible]			1		1									4					1					1					171
N																				10				1	1										1			2												18
l																					59						1				1				1	1														75
r																						91													1								1							103
w						1						1	2										35								1								1		1									44
j																								14																										14
i																									74	8							1					2												88
I												1			1	1									5	79						1	4		6		1	2												117
E																										2	62							3	4	1					1		3		1					83
{																											1	46	1												2									51
A																											1		40																					41
Q																														10											1									13
O																					3										30																			37
U																																3																		3
u																										5							40		2						1									55
3																																		3																3
@		1			1	1						2	1	1	1			1			1	1			1	2	3							1	84	1		1			1	1								150
V																				1									1	1					1	29					1									35
eI																												1								2	44													51
aI																									3												1	44					1							49
oI																																							1											1
aU																																					2			14										16
@U																										3					1			3	2			2			40		1			1				36
I@																																										16	1							17
e@																											1																1							2
U@																																												0						0
<ax>			1											1	3				1							4	1								6										4					21
<ah>						1																				1															1					3				6
<af>				1		2																													2						1						5			11
<an>																														1																		1		2
<ux>																																																	0	0

Table C13. Confusion matrix for Speaker 1

Appendix C. Results in full

	s	S	z	Z	f	v	T	D	h	tS	dZ	p	b	t	d	k	g	m	n	N	l	r	w	j
s	118		9																					
S		31																						
z	6		41																5					
Z				0																				
f					70		1					2				1								
v						25							1		1									
T					1		23					2	1	1										
D								0							1									
h							1		10															
tS										8	1													
dZ										1	9													
p												37			2									
b													9										1	
t	2		1		1		2					5		197	4	1								
d						1						2	2	4	52				2		2			
k														1	1	90								
g															2	4	32							
m																		42	5					
n															1			8	137		1			
N																			2	9				
l																					65		1	
r																						99		
w						3										1							38	
j																								10
i																								3
I			1		1										1				3					1
E															2									
{																			1					
A																								
Q																					1			
O																								
U																								
u																		1	1					
3																								
@															2	1			1		1			
V																								
eI																								
aI																								
oI																								
aU																								
@U																					1			
I@																								
e@																								
U@																								
[illegible]																			1					
[illegible]																								
[illegible]																			1					
[illegible]																								
[illegible]																								

	i	I	E	{	A	Q	O	U	u	3	@	V	eI	aI	oI	aU	@U	I@	e@	U@	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	Total
s																										136
S																										31
z		1																								57
Z																										0
f																										76
v											1															30
T																										35
D																										1
h																										13
tS																										9
dZ											1															11
p																										41
b																										15
t																										231
d											1															82
k																										97
g													1													41
m																										49
n		1									3	1														171
N	1																	1								18
l			1					2			1	1					1									75
r													1													103
w																										44
j	1	1																								14
i	74	4							4		1															88
I	3	83		1					3		3			2									1		1	117
E			70	1							4			1												83
{			2	51	1						1					1										57
A					35																					35
Q						29																				30
O							35																	1		37
U								2																		3
u									46		4				1		1									55
3										3																3
@		5	4						3	2	99	3					1	1			1		1	1		150
V			1		1	1					2	9					1									18
eI					2							1	48													51
aI														49												49
oI															1											1
aU																14										15
@U								1		1	1	3	1				46									1756
I@																		17								17
e@			1																1							2
U@																				0						0
[illegible]		1									8						2				8					21
[illegible]										1							3					2				6
[illegible]			1								4	1									1		3			11
[illegible]							1																	1		2
[illegible]																									0	0

Table C14. Confusion matrix for Speaker 2

Appendix C. Results in full

	s	S	z	Z	f	v	T	D	h	tS	dZ	p	b	t	d	k	g	m	n	N	l	r	w	j	i	I	E	{	A	Q	O	U	u	3	@	V	eI	aI	oI	aU	@U	I@	e@	U@	<ax>	<eh>	<af>	<or>	<n>	Totals
s	118		3				1			1			4																			1																136
S		30																																																	31
z	4		47																																																57
Z				0																																															0
f					65	1	4						1		3	1																																				76
v					1	15																													1																30
T					3		27	1								2																																			36
D								0																																												0
h	1								8			2																																								13
tS		1								8																																									9
dZ										1	7				1																																			11
p					1				1			34		1		1			1																																41
b													1	8										2	1														1																15
t										2		2		201	15				1									1										3			1							1					231
d	1					1			1	1				6	58				1														1		1																82
k	2						1								1	88	1																																	97
g														1		1	36				2																														41
m																		45	1																1																49
n																		11	132		1														8			1												171
N																		1	5	9																														18
l														1					1		57	1	3													1	1															75
r										1									1		1	91														3																103
w																		1				1	40								1																					44
j																								12		1													1														14
i																									45	2																									46
I																									19	122	2								3			1														160
E																											71								1	1								2							83
{																											4	49									1			1	1										57
A																					1						1		33																						35
Q																														11																					13
O																															37																				37
U																																2																			3
u						1																1		1			5								36		3								3					1				55
3																																		3																	3
@						1						1			1			1	1		1		1		1	6	6	1			1			3	97						2									150
V																		2								1										1	26	2																35
eI																													1								48			2												51
aI																									2	2														45													49
oI																																							1													1
aU																												1												15											16
@U																							1			1	1								5	2			1		43					1				56
I@																									1	1										1								14								17
e@																											1																			1							2
U@																																												0						0
<ax>														2	1											2							2	5												9					21
<eh>																																			1						2					3				6
<af>						1															1					1															1						5			11
<or>																					1										1																	0		2
<n>																																																	0	0

Table C15. Confusion matrix for Speaker 3.

Appendix C. Results in full

		Total
s	350 23 2 1 1 7 1	406
S	87 1	93
z	13 137 3 5 1	171
Z	0	0
f	205 1 7 5 3 1 1	234
v	1 51 2 1 1	90
T	1 4 80 1 2 1 1 2 6 1	107
D	0 1	1
h	1 1 26 1 2 1 1	39
tS	3 22 1 1	27
dZ	1 2 24 1 1	33
p	2 2 109 1 2 1 1 1	123
b	2 26 2 2 2	45
t	5 3 1 2 7 593 23 2 1 1 1 1 1 1 2	603
d	1 2 1 1 1 2 2 15 159 3 2 1 1 1 4 1	246
k	1 1 2 1 2 265 1 1	291
g	1 1 3 9 99 2 1 1	123
m	2 5 106 13 1 1 1	147
n	7 8 31 363 2 9 1 2 15 1 1 1 1	513
N	1 7 28 1 2 1 2 1	54
l	1 1 101 1 4 2 1 2 3 3 1	225
r	1 3 1 281 4 1 1	309
w	4 1 2 1 1 1 113 2 1 1	132
j	36 1 2 1	42
i	1 1 3 193 14 5 1 2	224
I	1 2 1 3 1 29 284 2 1 1 7 14 1 5 1 1 1	394
E	2 2 203 1 3 9 2 1 1 5	249
{	1 7 146 2 1 1 2 3	165
A	1 2 108	111
Q	1 50 1	56
O	3 102 1	111
U	1 7	9
u	1 1 1 1 1 10 124 9 1 5 1	165
3	9	9
@	1 1 2 3 1 1 4 1 2 2 3 1 1 2 13 13 1 1 3 6 280 4 1 4 2 1 1 1	450
V	2 1 1 1 2 2 4 64 2 2	88
eI	1 3 3 140 2	153
aI	5 2 1 138 1	147
oI	3	3
aU	1 2 43	48
@U	1 1 1 4 1 1 1 4 8 5 1 2 1 129 1 2	168
I@	1 1 1 47 1	51
e@	3 3	6
U@	0	0
<sil>	1 3 4 2 7 2 19 2 21	63
<ah>	1 1 1 1 6 8	18
<uf>	1 5 1 1 1 1 6 1 2 1 13	33
<um>	1 1 1 2 2	6
<un>	0	0

Table C16. Confusion matrix for all speakers.

Appendix C. Results in full

		%Recognised							
		Plo	Aff	SF	WF	L/G	Nas	Vow	Total
	Plosive	84.8	0.4	1.4	0.4	0.2	.	1.4	507
S	Affricate	5.0	70.0	15.0	.	.	5.0	.	20
p	Str Fric	2.7	0.9	90.6	0.4	.	.	.	224
o	Wk Fric	3.8	0.6	0.6	77.1	.	.	2.5	157
k	Liq/Glide	1.3	.	.	0.4	84.3	.	3.8	236
e	Nasal	8.0	.	.	0.8	4.6	60.5	4.2	238
n	Vowel	0.9	.	0.1	0.2	0.6	0.2	87.2	868

Table C17. Confusion matrix for manner of articulation – Speaker 1.

		%Recognised							
		Plo	Aff	SF	WF	L/G	Nas	Vow	Total
	Plosive	87.8	.	0.6	0.8	0.6	0.4	0.4	507
S	Affricate	.	95.0	.	.	.	.	5.0	20
p	Str Fric	.	.	91.5	.	.	2.2	0.4	224
o	Wk Fric	6.4	.	.	83.4	.	.	0.6	157
k	Liq/Glide	0.4	.	.	1.3	90.3	.	3.8	236
e	Nasal	0.4	.	.	.	0.4	85.3	2.9	238
n	Vowel	0.7	.	0.1	0.1	0.5	0.8	91.0	868

Table C18. Confusion matrix for manner of articulation – Speaker 2.

		%Recognised							
		Plo	Aff	SF	WF	L/G	Nas	Vow	Total
	Plosive	89.7	0.6	0.6	1.0	1.0	0.6	1.2	507
S	Affricate	5.0	80.0	5.0	.	.	.	.	20
p	Str Fric	1.8	0.4	90.2	0.4	.	.	0.4	224
o	Wk Fric	5.7	.	0.6	79.6	.	.	0.6	157
k	Liq/Glide	0.4	0.4	.	.	87.3	1.3	3.4	236
e	Nasal	.	.	.	.	0.4	85.7	4.2	238
n	Vowel	0.2	.	.	0.2	0.8	0.5	91.8	868

Table C19. Confusion matrix for manner of articulation – Speaker 3.

		%Recognised							
		Plo	Aff	SF	WF	L/G	Nas	Vow	Total
	Plosive	87.4	0.3	0.8	0.7	0.6	0.3	1.0	1521
S	Affricate	3.3	81.7	6.7	.	.	1.7	1.7	60
p	Str Fric	1.5	0.4	90.8	0.3	.	0.7	0.3	672
o	Wk Fric	5.3	0.2	0.4	80.0	.	.	1.3	471
k	Liq/Glide	0.7	0.1	.	0.6	87.3	0.4	3.7	708
e	Nasal	2.2	.	.	0.3	1.8	77.2	3.8	714
n	Vowel	0.6	.	0.1	0.2	0.6	0.5	89.9	2607

Table C20. Confusion matrix for manner of articulation – all speakers

Appendix C. Results in full

						%Recognised						
		Lab	Alv	P–A	Vel	Frn	Bck	Cen	F'g	B'g	C'g	Total
	Labial	82.3	0.3	0.3	0.3	.	0.3	1.7	0.3	0.3	.	293
	Alveolar	0.9	82.2	0.2	0.4	0.5	0.2	1.1	0.1	0.1	0.1	855
S	Pal–Alv	.	3.1	90.8	.	.	.	.	.	.	.	65
p	Velar	1.8	2.4	1.8	82.8	0.6	.	0.6	1.2	.	.	169
o	Front	.	.	.	.	81.7	2.1	4.1	1.5	0.9	0.9	339
k	Back	.	0.7	.	.	4.0	82.6	1.3	.	1.3	.	149
e	Central	3.2	1.6	0.5	0.5	3.2	1.1	63.3	0.5	1.1	0.5	188
n	Fronting	.	.	.	.	4.0	.	2.0	89.1	.	1.0	101
	Backing	.	.	.	.	4.2	1.4	6.9	5.6	75.0	1.4	72
	Centring	.	.	.	.	5.3	.	.	.	.	94.7	19

Table C21. Confusion matrix for place of articulation – Speaker 1.

						%Recognised						
		Lab	Alv	P–A	Vel	Frn	Bck	Cen	F'g	B'g	C'g	Total
	Labial	87.4	3.4	.	0.7	.	.	0.3	.	.	.	293
	Alveolar	2.6	86.7	.	0.1	0.4	0.2	0.8	0.1	0.1	.	855
S	Pal–Alv	.	.	92.3	.	3.1	.	1.5	.	.	.	65
p	Velar	0.6	3.6	.	85.7	0.6	.	.	0.6	.	0.6	169
o	Front	0.3	2.3	1.2	.	84.3	2.3	3.2	0.9	0.3	.	345
k	Back	0.6	1.3	.	.	.	91.9	.	0.6	0.6	.	160
e	Central	.	2.9	.	.	5.8	2.9	69.0	.	1.2	0.6	171
n	Fronting	.	.	.	.	.	2.0	1.0	97.0	.	.	101
	Backing	.	1.4	.	.	.	1.4	6.9	1.4	83.3	.	72
	Centring	.	.	.	.	5.3	.	.	.	.	94.7	19

Table C22. Confusion matrix for place of articulation – Speaker 2.

						%Recognised						
		Lab	Alv	P–A	Vel	Frn	Bck	Cen	F'g	B'g	C'g	Total
	Labial	88.7	4.1	.	0.3	.	0.3	1.0	.	.	.	293
	Alveolar	2.1	87.1	0.6	0.1	0.1	0.1	1.9	0.2	.	.	855
S	Pal–Alv	.	.	90.8	.	1.5	.	.	1.5	.	.	65
p	Velar	2.4	7.1	.	84.6	.	.	.	.	.	.	169
o	Front	.	.	.	.	90.2	.	1.4	0.6	0.6	0.6	348
k	Back	1.4	1.4	.	.	4.2	84.6	2.1	.	2.1	.	143
e	Central	3.2	1.6	.	.	7.8	0.5	69.1	1.1	1.1	.	188
n	Fronting	.	.	.	.	5.0	.	.	93.1	2.0	.	101
	Backing	.	.	.	.	4.2	.	9.7	1.4	80.6	.	72
	Centring	.	.	.	.	15.8	.	5.3	.	.	78.9	19

Table C23. Confusion matrix for place of articulation – Speaker 3.

						%Recognised						
		Lab	Alv	P–A	Vel	Frn	Bck	Cen	F'g	B'g	C'g	Total
	Labial	86.1	2.6	0.1	0.5	.	0.2	1.0	0.1	0.1	.	879
	Alveolar	1.9	85.4	0.3	0.2	0.3	0.2	1.2	0.2	0.1	.	2565
S	Pal–Alv	.	1.0	91.3	.	1.5	.	0.5	0.5	.	.	195
p	Velar	1.6	4.3	0.6	84.4	0.4	.	0.2	0.6	.	0.2	5.7
o	Front	0.1	0.8	0.4	.	85.5	1.5	2.9	1.0	0.6	0.5	1032
k	Back	0.6	1.1	.	.	2.7	86.5	1.1	0.2	1.3	.	547
e	Central	2.2	2.0	0.2	0.2	5.7	1.5	67.1	0.5	1.1	0.4	452
n	Fronting	.	.	.	.	3.0	0.7	1.0	93.1	0.7	0.3	303
	Backing	.	0.5	.	.	2.8	0.9	7.9	7.8	79.6	0.5	216
	Centring	.	.	.	.	8.8	.	1.8	.	.	89.5	57

Table C24. Confusion matrix for place of articulation – all speakers.

REPORT DOCUMENTATION PAGE

DRIC Reference Number (if known) ....

Overall security classification of sheet ....UNCLASSIFIED..

(As far as possible this sheet should contain only unclassified information. If it is necessary to enter classified information, the field concerned must be marked to indicate the classification eg (R), (C) or (S).

Originators Reference/Report No.	Month	Year
MEMO 4463	MARCH	1991

Originators Name and Location
RSRE, St Andrews Road
Malvern, Worcs WR14 3PS

Monitoring Agency Name and Location

Title

ANALYSIS OF THE PHONEME RECOGNITION PERFORMANCE OF THE ARM CONTINUOUS SPEECH RECOGNITION SYSTEM

Report Security Classification	Title Classification (U, R, C or S)
UNCLASSIFIED	U

Foreign Language Title (in the case of translations)

Conference Details

Agency Reference	Contract Number and Period

Project Number	Other References

Authors	Pagination and Ref
BROWNING, S R	30

Abstract

This memorandum presents the results of a phonetically motivated analysis of the speech recognition system developed as part of the *ARM* (Airborne Reconnaissance Mission) project. The aim of the work described here is to investigate to what extent errors can be explained by phonetic effects; those which cannot may indicate where models may be improved. The background to the investigation, and the problems of evaluating phoneme recognition performance are described, then the remainder of the report is concerned with a detailed analysis of specific types of errors, motivated by a desire to find phonetic explanations of them.

Abstract Classification (U,R,C or S)
U

Descriptors

Distribution Statement (Enter any limitations on the distribution of the document)

UNLIMITED

580/48

INTENTIONALLY BLANK